Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 5 | May 2017

WEB MINING AGAINST PEDOPHILIA

Shweta Macwan [1] | Dr. inż. Grzegorz Filcek [2]

[1] Student, Information Technology, Wroclaw University of Science and Technology, Wroclaw, Poland – 50-370.

[2] Assistant Professor, Information Technology, Wroclaw University of Science and Technology, Wroclaw, Poland – 50-370.

ABSTRACT

The need of security over the web is the foremost necessity and handling the cybercrimes is a priority. The growing popularity of the social media has led the children to use the internet more for social communication than information gathering. Children needs to learn and grow with technology but child safety is also required. Pedophiles hunt for innocent children over such social media and chat room platforms which are not safe for the child. Due to lack of parental guidance, such cases lead to cybercrimes which kids are not aware of. Social media is not the only area where pedophilic activities takes place. The search on the search engine may also help in detecting a pedophile. Here, the main idea is to capture the pedophiles using the conversions made with a child and detecting it based on the pattern of words and language used by an adult. Also, with the help of the search engine's query detection a pedophilic activity can be traced.

KEYWORDS: Web mining, Web content mining, Pedophile, cyber-crime, cyberpedophilia, pedophilic activity.

INTRODUCTION

Protection of children on cyber space is an extremely critical problem faced by our society across geographical and cultural boundaries. As more and more children in their teens have started using the Internet, there has been an alarming increase in cases of child abuse through the Web.[1] As a report published by the National Center for Missing and Exploited Children (NCMEC),1 out of 7 kids is solicited for sex online; 1 out of 33 kids receives aggressive online solicitation to meet in person 1 out of 3 kids receives unsolicited sexual content online.[2] Internet nowadays is providing an easy and convenient access to the predators or criminals. Parents on the other hand does not track how their children are using the Internet. The lack of attention from parents and the criminal intentions of some people gives birth to cybercrime in children. The pedophiles are people having a psychic disorder and are sexually attracted towards the prepubescents. Today cybercrime activities such as pedophilia activities are a major issue of concern. This activity is termed a cyberpedophilia. Children are not safe on the internet and there is a dire need for an internet space that is safe for children. It has always been recommended that parents monitor their children's activities on the internet as to what they post, what they see, whom do they chat to, what kind of messages so they receive.

The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet. With the Internet usage gaining popularity and the steady growth of users, the World Wide Web has become a huge repository of data and serves as an important platform for the dissemination of information. Web mining can then be defined as for the discovery and analysis of useful information from the World Wide Web. The combination of Data Mining and World Wide Web is termed as Web Mining. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining is the application of data mining technique to web data to discover useful patterns. The data available on web is termed as Web data and the process of mining the web data is termed as Web mining. The most commonly used techniques are association rule, classification, clustering and sequential pattern identification.

Web data are usually in the following forms: Web content that includes text, images, structured records, videos, audio files etc., Web Structure that includes hyperlinks, document structure and tags and Web Usage that includes web server logs, application server logs and application level logs.

For any type of mining the most important step to be done is preprocessing. Data preprocessing is the step where the raw data is processed in such a way that the extracted data would be useful to mine some knowledge. There are some levels of processing done on the raw data to obtain a knowledgeable data. These levels include selecting the target data from the raw data, extraction of some data and transform the processed data to obtain knowledge. The processing includes cleaning of noisy data, integrating the data, data transformation, data reduction and data discretization.

There are three types of web mining techniques based upon the usage and the type of knowledge to be mined and extracted. Web usage mining, web structure mining and web content mining.

Taxonomy of web mining:

A. Web Structure Mining

Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. The preprocessing of this type of mining involves identifying interesting graph patterns or preprocessing the whole web graph to come up with some matrices. The most common example is PageRank. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting between two related pages. Such mining can be done on intra-page document level or inter-page hyperlink level. Some of the major use of this technique is done in PageRank algorithm, Hubs and Authorities, HITS algorithm, Information Scent, etc. Useful information such as quality of a web page, interesting web structures and web page classification can be obtained.

B. Web Usage Mining

Web usage mining is the process of applying data mining techniques to the discovery of usage patterns from web data. The data available on the web is not only huge but also semi-structured. The browsing history of the user is stored in a log file which can be used to mine interesting patterns. logs, proxy logs or browser logs. These log files hold a lot of information such as URLs, IP addresses, time, date, etc. When people visit one website, they leave some data such as IP address, visiting pages, visiting time and so on, web usage mining will collect, analyses and process the log and recording data. [3] This technique is widely used in ecommerce, web transactions, path and pattern discovery, pattern analysis and many more.

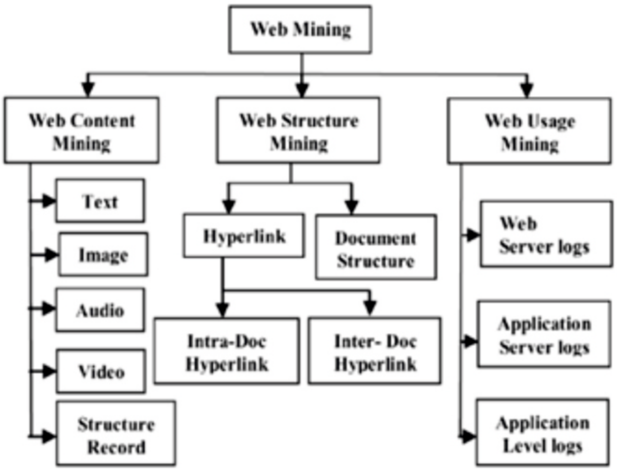


Fig. 1 Taxonomy of Web Mining

C. Web Content Mining

Web content mining is the process to discover useful information from text, image, audio or video data on the web. Information retrieval is the basic means for any information gathering technique which helps user to find the specific information from the large set of data.[4] This technique is mainly used for Natural

Copyright© 2016, IERJ. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

Language Processing and Information Retrieval. Due to enormous size, a web query can result in multiple results possibly with repetition. Thus, we need to present a technique that reduces the amount of information in the result that is appropriate for our knowledge mining.

Web search deals with Information retrieval, a study that helps to retrieve knowledge and information from a vast dataset most likely to be web data in this case.

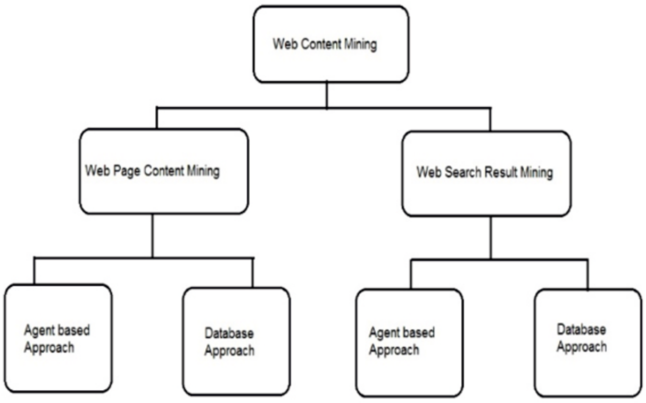


Fig. 2 Web Content Mining

The efforts can be grouped into two subcategories:

(i) Agent based Approach: The agent approach uses so called Web agents to collect relevant information from the World Wide Web. A Web agent is a program that visits a Web site and filters the information the user is interested in. There are three subtypes for the agent based approach: Intelligent Search Agents, Information Filtering/Categorization and the Personalized Web Agents. For more information about these subtypes.[5]

(ii) Database Approach: The database approach for Web mining tries to develop techniques for organizing semi structured data stored in the Web into more structured collections of information resources. Standard database querying mechanisms and data mining techniques can be used to analyze those collections then.[5]

Preprocessing of Web Data:

The content data needs to be preprocessed before the actual knowledge mining process. The content is in an unstructured form and has many components that are not useful or important for our knowledge.

Steps for preprocessing the content of Web data are as follows:

1. Extract text from HTML

Any web data is in the form of HTML page. The aim is to extract the data that are available on the web. This data is in the form of text, audio, video, animations etc.

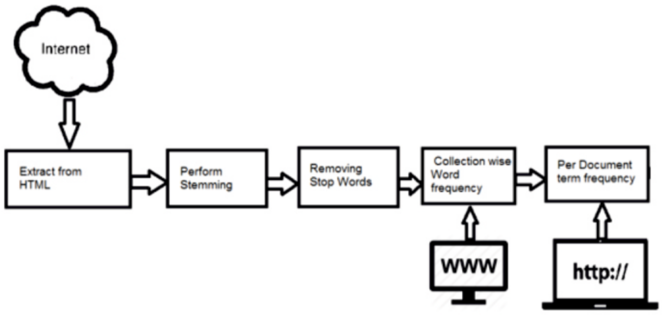


Fig 3. Content Preparation

The required data that is useful for the mining process needs to be extracted so that processing can be done and useful information can be obtained from the extracted data. The aim is to decide what type of data is required for our knowledge and how much data needs to be extracted from the whole web data.

2. Perform Stemming

Stemming is the process of deducing words from the text. The process includes reducing the words from the word-stem. Many search engines treat word and word-stem as synonyms to increase the result of search query. Words such as "fishing", "fished" are "fisher" stem words for the root word "fish".

3. Removing Stop Words

Stop words are words that are most commonly in the language. These words are most widely used for natural language processing tools. There is no group of words for the stop words. Words such as "I", "you", "the", "it", "what", "as" etc. are some of the common stop words.

4. Calculate Collection Wise Word Frequencies (DF)

The text collectively has repetitive words which are to be removed from the text. The text here is considered as the whole website which includes all the webpages linked to each other. Thus, unique words are identified and extracted from the web data. These unique words are also repetitively used in the text therefore the frequency of each unique word is calculated.

5. Calculate per Document term frequency(TF)

Here, the text is considered only for the single webpage. The frequency of each unique word is calculated from the web page.

The objective is to preprocess the chat log data using the data mining preprocessing technique. The processed chat log helps to distinguish the users in the conversation. Based on the terminology, language of an individual, usage of negative words and its frequency, identification of pedophile is done. Also, using the negative words as a base, detection of a pedophilic activity can be done on the search engine.

The advancement of Information and Communication Technology has led to various innovations in our personal lives. Today chat servers are a vital part in life. Messaging applications such as Facebook, Instant Messenger, Yahoo Messenger, WhatsApp, etc. are most popular amongst the youth nowadays. However, these messaging applications cannot be controlled or managed. The peer-to-peer chat conversation is almost impossible to keep a track on. The main issue with such messaging system is that it is difficult to know the person on the opposite side. The person on the other side can be a predator for the young children. This is only possible to know if we have the conversation and the way the predator builds a relationship with the child.

Apart from conversations, predators can be tracked down with the help of search engines. Predators tend to search for images, videos and other related content to fulfill their needs. Such activity should also be controlled and verified. With the enormous amount of data available on the internet, pedophiles can be more active and get more information than needed. Such data should be restricted that are not safe for a child's security. Such major issues need to be controlled and the pedophiles should be kept isolated from such data.

Experiments:

The problem deals with many aspects in the field of web data related to pedophilia activities. Here, the focus is mainly on the parts related to search systems and chat logs.

A. Chat Logs

The chat servers are widely used in today's day to day life. The data obtained from chat logs contain plenty of information. Text mining is to be applied on the data available as chat data. There is an American foundation Perverted Justice(PJ) where who investigates cases of online child sexual abuse. Adult volunteers enter chat rooms as juveniles (usually 12-15-year-old) and if they are sexually solicited by adults, they work with the police to prosecute the offenders. Some chat conversations with cyberpedophiles are available at www.perverted-justice.com and they have been the subject of analysis of recent research on this topic. [6]

The chat lines are mainly categorized into the following: [7]

1. Exchange of personal Information
2. Grooming
3. Approach
4. None of the above classes [8]

Step 1: Exchange of personal information
Pedophile: Hey beautiful
Pedophile: What's your number

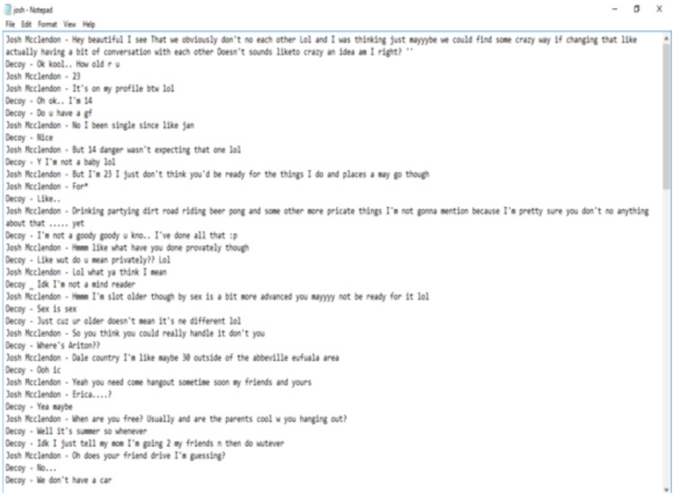
Step 2: Grooming
Pedophile: Yeah you need come hangout sometime soon my friends and yours
Pedophile - Hmmm like what have you done privately

Step 3: Approach
Predator: licking don't hurt
Predator: it's like u lick ice cream
Pseudo-victim: do u care that I'm 13 in March and not yet? I lied a little bit b4
Predator: it's all cool

Step 4: None of the above
Predator: don't tell anyone we have been talking
Pseudo-victim: k
Pseudo-victim: lol who would I tell? No one's here.
Predator: well I want it to be our secret

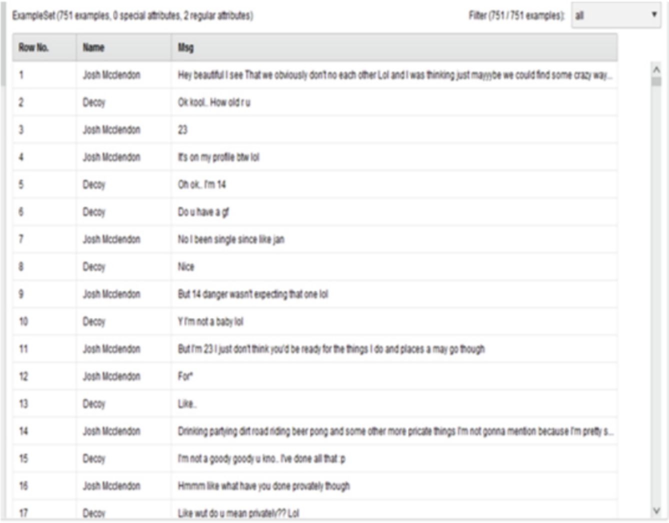
The idea is to notify that a person is a pedophile or not based on the language and writing methods used by a person during an online chat conversation. The chat logs are taken from the perverted justice website where the data is available for research purposes.

The figure 4 shows the chat log between Josh and Decoy that was available on the PJ website. This data needs to be processed to gain knowledge. The steps categorization of the chat can be seen. The data then undergoes preprocessing because of the humongous amount of information and raw data. The chat log data is an unstructured for that needs to be processed and formatted into a structured form of data which will help to easily process the data as per the requirement or need of the system.



```
Josh Mcclendon - Hey beautiful I see That we obviously don't no each other Lol and I was thinking just mayyybe we could find some crazy way if changing that like
actually having a bit of conversation with each other Doesn't sounds liketo crazy an idea am I right? ''
Decoy - Ok kool.. How old r u
Josh Mcclendon - 23
Josh Mcclendon - It's on my profile btw lol
Decoy - Oh ok.. I'm 14
Decoy - Do u have a gf
Josh Mcclendon - No I been single since like jan
Decoy - Nice
Josh Mcclendon - But 14 danger wasn't expecting that one lol
Decoy - Y I'm not a baby lol
Josh Mcclendon - But I'm 23 I just don't think you'd be ready for the things I do and places a may go though
Josh Mcclendon - For*
Decoy - Like..
Josh Mcclendon - Drinking partying dirt road riding beer pong and some other more pricate things I'm not gonna mention because I'm pretty sure you don't no anything
about that ..... yet
Decoy - I'm not a goody goody u kno.. I've done all that :p
Josh Mcclendon - Hmmm like what have you done provately though
Decoy - Like wut do u mean privately?? Lol
Josh Mcclendon - Lol what ya think I mean
Decoy _ Idk I'm not a mind reader
Josh Mcclendon - Hmmm I'm slot older though by sex is a bit more advanced you mayyyy not be ready for it lol
Decoy - Sex is sex
Decoy - Just cuz ur older doesn't mean it's ne different lol
Josh Mcclendon - So you think you could really handle it don't you
Decoy - Where's Ariton??
Josh Mcclendon - Dale country I'm like maybe 30 outside of the abbeville eufuala area
Decoy - Ooh ic
Josh Mcclendon - Yeah you need come hangout sometime soon my friends and yours
Josh Mcclendon - Erica....?
Decoy - Yea maybe
Josh Mcclendon - When are you free? Usually and are the parents cool w you hanging out?
Decoy - Well it's summer so whenever
Decoy - Idk I just tell my mom I'm going 2 my friends n then do wutever
Josh Mcclendon - Oh does your friend drive I'm guessing?
Decoy - No...
Decoy - We don't have a car
```

Fig 4. Chat Log

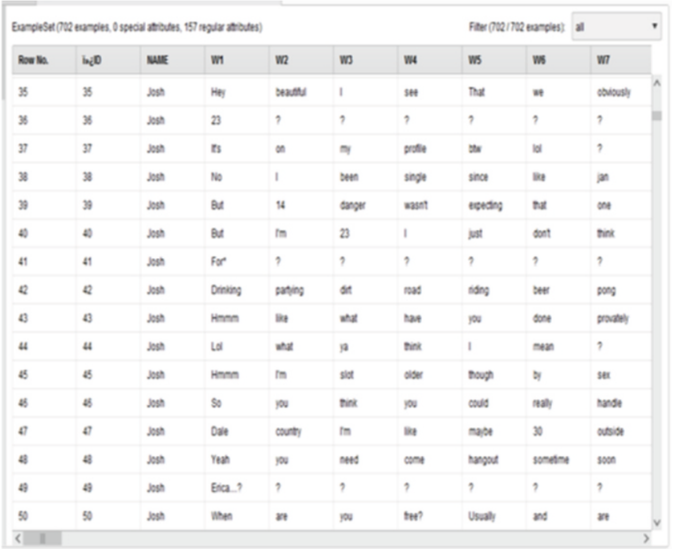


ExampleSet (751 examples, 0 special attributes, 2 regular attributes) Filter (751 / 751 examples): all

Row No.	Name	Msg
1	Josh Mcclendon	Hey beautiful I see That we obviously don't no each other Lol and I was thinking just mayyybe we could find some crazy way...
2	Decoy	Ok kool.. How old r u
3	Josh Mcclendon	23
4	Josh Mcclendon	It's on my profile btw lol
5	Decoy	Oh ok.. I'm 14
6	Decoy	Do u have a gf
7	Josh Mcclendon	No I been single since like jan
8	Decoy	Nice
9	Josh Mcclendon	But 14 danger wasn't expecting that one lol
10	Decoy	Y I'm not a baby lol
11	Josh Mcclendon	But I'm 23 I just don't think you'd be ready for the things I do and places a may go though
12	Josh Mcclendon	For*
13	Decoy	Like..
14	Josh Mcclendon	Drinking partying dirt road riding beer pong and some other more pricate things I'm not gonna mention because I'm pretty s...
15	Decoy	I'm not a goody goody u kno.. I've done all that :p
16	Josh Mcclendon	Hmmm like what have you done provately though
17	Decoy	Like wut do u mean privately?? Lol

Fig 5. Text Extraction

Fig. 5 shows the data of chat log into a structured format which will be further needed to for processing of this data. The idea is to differentiate the raw data and manage it in a meaningful manner that is easily understandable and usable for the mining process.



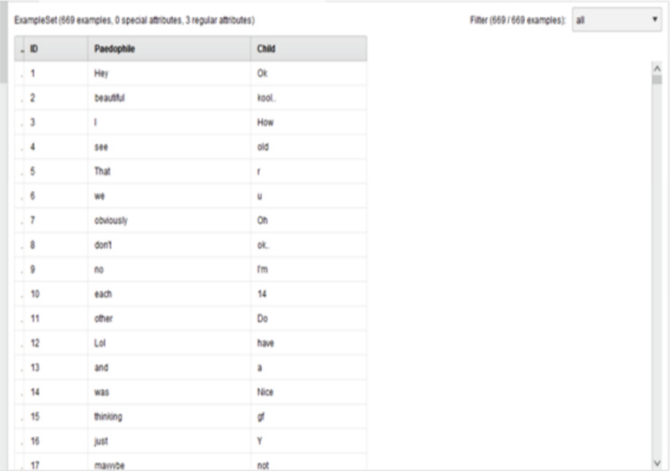
ExampleSet (702 examples, 0 special attributes, 157 regular attributes) Filter (702 / 702 examples): all

Row No.	logID	NAME	W1	W2	W3	W4	W5	W6	W7
35	35	Josh	Hey	beautiful	I	see	That	we	obviously
36	36	Josh	23	?	?	?	?	?	?
37	37	Josh	It's	on	my	profile	btw	lol	?
38	38	Josh	No	I	been	single	since	like	jan
39	39	Josh	But	14	danger	wasn't	expecting	that	one
40	40	Josh	But	I'm	23	I	just	don't	think
41	41	Josh	For*	?	?	?	?	?	?
42	42	Josh	Drinking	partying	dirt	road	riding	beer	pong
43	43	Josh	Hmmm	like	what	have	you	done	provately
44	44	Josh	Lol	what	ya	think	I	mean	?
45	45	Josh	Hmmm	I'm	slot	older	though	by	sex
46	46	Josh	So	you	think	you	could	really	handle
47	47	Josh	Dale	country	I'm	like	maybe	30	outside
48	48	Josh	Yeah	you	need	come	hangout	sometime	soon
49	49	Josh	Erica...?	?	?	?	?	?	?
50	50	Josh	When	are	you	free?	Usually	and	are

Fig 6. Word Extraction

Fig. 6 shows the chat log conversation differentiating the whole text into individual words. These individual words are further used as information retrieval and mining process. This is a part of the preprocessing of the data that is unstructured to make it a simpler form that is used in the knowledge gaining process.

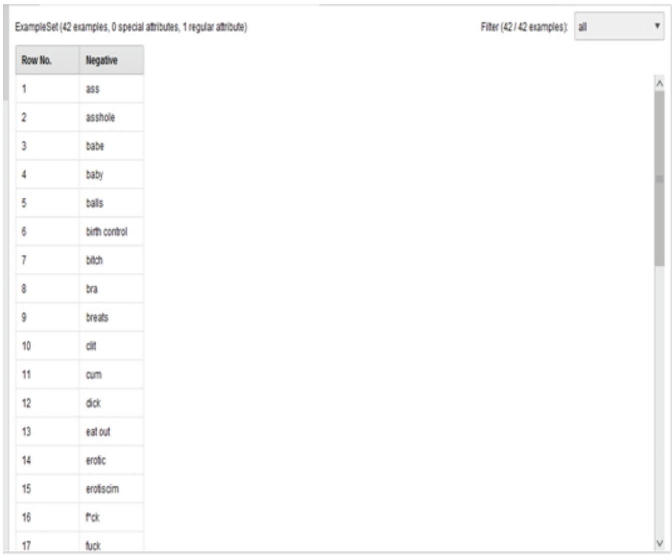
Fig.7 shows the list of distinct words that are used in the conversation between Josh and Decoy.



ExampleSet (669 examples, 0 special attributes, 3 regular attributes) Filter (669 / 669 examples): all

ID	Paedophile	Child
1	Hey	Ok
2	beautiful	kool..
3	I	How
4	see	old
5	That	r
6	we	u
7	obviously	Oh
8	don't	ok..
9	no	I'm
10	each	14
11	other	Do
12	Lol	have
13	and	a
14	was	Nice
15	thinking	gf
16	just	Y
17	mayybe	not

Fig 7. Distinct Word Extraction

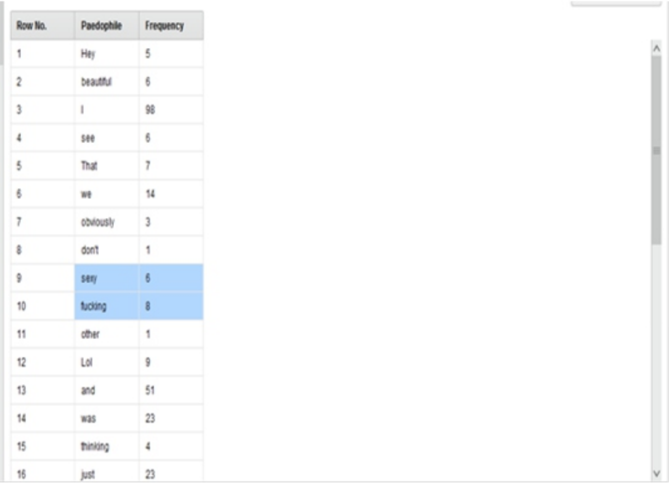


ExampleSet (42 examples, 0 special attributes, 1 regular attribute) Filter (42 / 42 examples): all

Row No.	Negative
1	ass
2	asshole
3	babe
4	baby
5	balls
6	birth control
7	bitch
8	bra
9	breats
10	clit
11	cum
12	dick
13	eat out
14	erotic
15	erotiscim
16	f*ck
17	fuck

Fig. 8 Negative Words

The fig. 9 shows the frequency of the words in the conversation between Decoy and Josh. As seen, the word "sexy" and "fucking" are repeated many times in the conversation. These words fall under the category of negative words and they play a vital role to decide whether the person is pedophile or not. Based on this record we can say that Josh is a pedophile and is a threat for Decoy a.k.a. Erica.



Row No.	Paedophile	Frequency
1	Hey	5
2	beautiful	6
3	I	98
4	see	6
5	That	7
6	we	14
7	obviously	3
8	don't	1
9	sexy	6
10	fucking	8
11	other	1
12	Lol	9
13	and	51
14	was	23
15	thinking	4
16	just	23

Fig. 9 Frequency of each word in the conversation

The frequency of each word is calculated that are distinctively used in the conversation. The frequency of words decides the level of pedophilic activity in the conversation. The negative set words are developed based on the pedophilic activities on the internet. The words in the conversations are compared to these set of negative words. The frequency of the negative words is higher than the chances of the person to be a pedophile increases as compared to the frequency of the positive words. The increase in the negative words describes the person is a pedophile and he/she is a threat to the child. This data can be tracked by the police or the cybercrime department and restrictions can be put against such person and activity.

B. Content Search

The search engine does not provide any restrictions on the search query. The information available online can be hazardous to many including children. The data that is accessible over the web is not secured and mainly does not deal with privacy or age verification. The data that is viewed by millions of end-users does not have to deal with the owner of the data. The end user searches for information over the web that contains data such as images and videos. If the data requested by the user is inappropriate, then limitations must be kept for such users. For example, if a person is searching for a "mountain", it does not deal with any of illegal activity or negative words. Therefore, that person can be considered safe and does not deal with any illegal activities. But if the person searches for an image of a naked child, he is a pedophile and response should not be available for his request as he is a threat for the society, especially children. Such data is not accessible for end user and only the cybercrime department has the authority to access such data. The search request prediction is not fully accurate because the data requested by the user might not contain the intention of the search. But assumptions can be made as to protect the child from being attacked by a pedophile. The request of any end user is approved by the search engine. Whereas, that should not be the scenario. All requests are not acceptable as some may lead to some illegal or criminal activities. Despite of the increase in cybercrimes, researchers are leading on the way as to how such crimes can be decreases or minimized.

The idea is to create a search page that restricts the user to search from the content that are related to pedophilia.

Fig. 11 shows that a user has request for details of a mountain. This word is not related to pedophile activities so the result would be displayed as required.

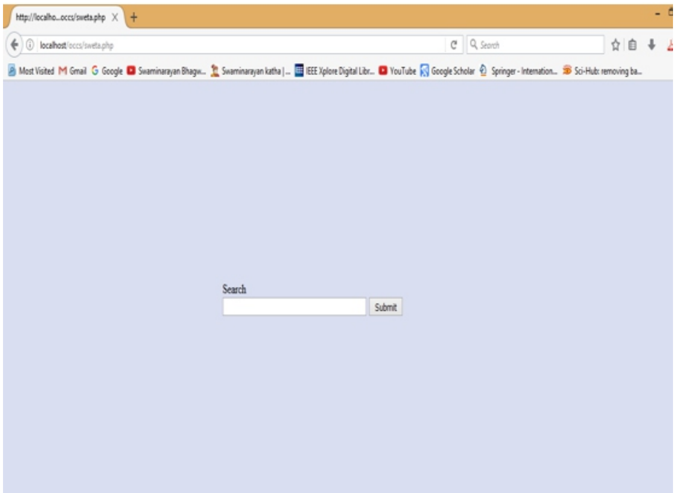
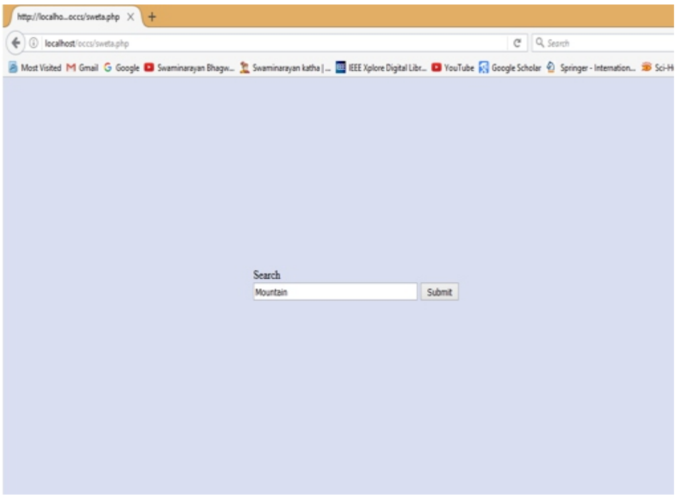
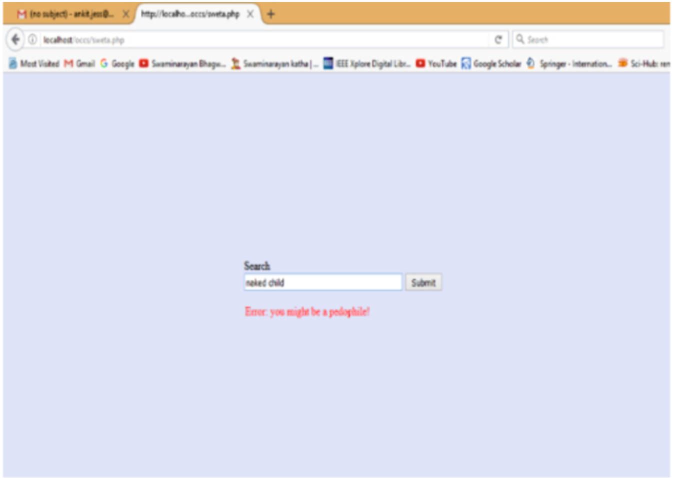


Fig. 10 Search Page



Mountain

Fig. 11 Search Page (1)



naked child

Error: you might be a pedophile!

Fig. 12 Search Page (2)

Fig. 12 shows the user has requested for details of a "naked child" and hence the details will not to be given to him/her because there is a threat that the person is a pedophile. Such restrictions are needed while developing a search engine systems.

CONCLUSIONS:

Creating a safe environment for children is the vital necessity. Children and teenagers have a door to link to the world with the internet. This work proposes various aspects in which pedophilic activities are possible and how they can be resolved using the chat logs and search engines. The chat systems propose to identify a pedophile. This may be used by the police or the cybercrime department to track the pedophile and stop such activities, keeping the children safe from predators.

Also, proposing a model for the search engines that restricts the usage and request queries from the user. The engine does not allow the user to obtain the data that the model might think would lead to a pedophilic activity. Thus, creating a safer internet-space for the children.

The chat model can be further extended and sentiment can be analyzed. Based on the sentiments, the level of pedophile can be obtained or predicted. The multilingualism constraint can be resolved and much accurate result can be obtained.

The search model leads to a restricted search. This restriction can be extended on various terms and areas and therefor leading limitations of search for many illegal criminal activities. A general model can be used with any search engine thus leading towards a restricted internet. The police can keep a track on those predators who tries to search items only related to pedophilic activities. Such pedophiles can be tracked down through IP addresses.

The work proposed are limited to two aspects over the internet. There are various other aspects that are needed to be kept safe for children. The online video calling does not restrict the pedophile to socialize with the children. Due to lack of attention of parents, children fall in the trap of pedophile and sometimes even they are recorded.

REFERENCES:

[1] Gupta, Aditi, Ponnurangam Kumaraguru, and Ashish Sureka. "Characterizing pedophile conversations on the internet using online grooming." arXiv preprint arXiv:1208.4324 (2012).

[2] NCMEC, National center for missing and exploited children, 2008 http://www.missingkids.com/missingkids/servlet/NewsEventServlet?LanguageCountry=en US&PageId=4303.

[3] Pranit Bari and P.M. Chawan," Journal of Engineering, Computers & Applied Sciences" (JEC&AS) Volume 2, No.6, June 2013, ISSN No: 2319-5606

[4] Guandong Xu, Yanchun Zhang and Lin Li ,"Web Mining and Social Networking", pp 71-87, 2011, DOI 10.1007/978-1-4419-7735-9_4, ISBN:978-1-4419-7735-9, Springer US

[5] Patidar, Kamlesh, Preetesh Purohit, and Kapil Sharma. "Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library 1." (2011).

[6] Dasha Bogdanova, Paolo Rosso, Thamar Solorio, Exploring high-level features for detecting cyberpedophilia, Computer Speech & Language, Volume 28, Issue 1, January 2014, Pages 108-120, ISSN 0885-2308

[7] India McGhee, Jennifer Bayzick, April Kontostathis, Lynne Edwards, Alexandra McBride and Emma Jakubowski. Learning to identify Internet sexual predation.International Journal on Electronic Commerce,2011.

[8] Bogdanova, Dasha, Paolo Rosso, and Thamar Solorio. "On the impact of sentiment and emotion based features in detecting online sexual predators." Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Asso-

ciation for Computational Linguistics, 2012.

[9] Falcão Jr, Mário Sérgio Rodrigues, Enyo José Tavares Gonçalves, and Tciciana Linhares Coelho da Silva. "Behavioral Analysis for Child Protection in Social Network through Data Mining and Multiagent Systems." (2016).

[10] R. Cooley, B. Mobasher and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, 1997,pp.558-567.doi:10.1109/TAI.1997.

[11] Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona, Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it, Expert Systems with Applications, Volume 40, Issue 18, 15 December 2013, Pages 7478-7491, ISSN 0957-4174, http://dx.doi.org/10.1016/j.eswa.2013.07.040.

[12] Tarique Anwar, Muhammad Abulaish, A social graph based text mining framework for chat log investigation, Digital Investigation, Volume 11, Issue 4, December 2014, Pages 349-362, ISSN 1742-2876

[13] M. Ashcroft, L. Kaati and M. Meyer, "A Step Towards Detecting Online Grooming -- Identifying Adults Pretending to be Children," 2015 European Intelligence and Security Informatics Conference, Manchester, 2015, pp. 98-104.

[14] M. Munezero, M. Mozgovoy, T. Kakkonen, V. Klyuev and E. Sutinen, "Antisocial Behavior corpus for harmful language detection," 2013 Federated Conference on Computer Science and Information Systems, Krako??w, 2013, pp. 261-265.

[15] Munezero, M., Montero, C. S., Kakkonen, T., Sutinen, E., Mozgovoy, M., & Klyuev, V. (2014). Automatic detection of antisocial behaviour in texts. Informatica, 38(1), 3-10.

[16] Hofmann, Alfred, et al. "Detection of Child Sexual Abuse Media: Classification of the Associated Filenames."

[17] Matthieu Latapy, Clémence Magnien, Raphaël Fournier, Quantifying paedophile activity in a large P2P system, Information Processing & Management, Volume 49, Issue 1, January 2013, Pages 248-263, ISSN 0306-4573

[18] Y. Shavitt and N. Zilberman, "On the Presence of Child Sex Abuse in BitTorrent Networks," in IEEE Internet Computing, vol. 17, no. 3, pp. 60-66, May-June 2013.

[19] Mathiesen, Kay. "The Internet, children, and privacy: the case against parental monitoring." Ethics and Information Technology 15.4 (2013): 263-274.

[20] Abdelghani Guerbas, Omar Addam, Omar Zaarour, Mohamad Nagi, Ahmad Elhajj, Mick Ridley, Reda Alhajj, Effective web log mining and online navigational pattern prediction, Knowledge-Based Systems, Volume 49, September 2013, Pages 50-62, ISSN 0950-7051

[21] T. Anwar and M. Abulaish, "Ranking Radically Influential Web Forum Users," in IEEE Transactions on Information Forensics and Security, vol. 10, no. 6, pp. 1289-1298, June 2015.

[22] K. Sudheer Reddy, M. Kantha Reddy and V. Sitaramulu, "An effective data preprocessing method for Web Usage Mining," 2013 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, 2013, pp. 7-10.

[23] Westlake, Bryce, Martin Bouchard, and Richard Frank. "Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation." Sexual abuse: a journal of research and treatment (2015): 1079063215616818.

[24] Database: http://www.perverted-justice.com/